

# RAG における自己認識的不確実性の評価

二宮大空<sup>1</sup> 戸田隆道<sup>1</sup><sup>1</sup> 株式会社 AI Shift

{ninomiya\_hirotaka,toda\_takamichi}@cyberagent.co.jp

## 概要

大規模言語モデル (Large Language Model: LLM) の普及により、カスタマーサポート事業では外部知識に基づき回答する Retrieval Augmented Generation (RAG) の導入検討が積極的に行われている。しかし、ユーザーの質問に対する答えが検索で得られなかった場合、誤った情報を生成してしまう可能性が高い。このような場合、本来であれば回答不能であることを LLM が認識してユーザーに伝えることが望ましい。そこで、本研究では回答可能かに関する認識の正確さを自己認識的不確実性と定め、LLM を用いて定量的に評価する。さらに、LLM を用いた回答の修正が自己認識的不確実性の向上に有効であるかを検証する。

## 1 はじめに

LLM の普及に伴い、カスタマーサポートのチャットボットには高品質な対応が求められている。この分野では、ユーザーからの質問は製品やサービス情報など企業独自の情報に関連する内容が多い。しかし、LLM の学習には企業が独自に保有する情報は含まれていないことが多いため、単に LLM を使用するだけでは最新で誤りのない高品質な回答を生成することは難しい。

そこで、外部知識からユーザーの質問に関連する情報を抽出し、その情報に基づいた回答を生成することが可能な RAG[1, 2, 3] が注目されている。しかし、RAG はユーザーの質問に対する答えが検索により得られなかった場合、誤った回答を生成する可能性が高いという課題がある。このような場合、本来は「回答に十分な情報が見つかりませんでした」といったように回答不能であることをユーザーに伝えることが望ましい。特に、カスタマーサポートにおける社内ドキュメントは Web 上の文書などと比較すると小規模であるため、回答に必要な情報が検索対象となる外部知識に存在しない場合が多く、RAG

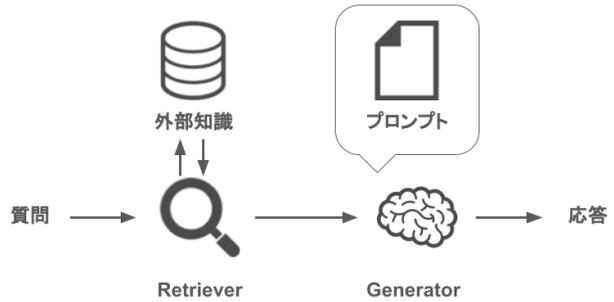


図 1 RAG の概要図

を事業応用における障壁となっている。そのため、RAG が回答可能かどうかを正しく認識できると、それを適切に評価できることが重要となる。

一般的に誤情報を生成する問題は hallucination[4]と呼ばれる。本研究で我々が注目する外部知識の不足による誤回答生成は、hallucination の中でも特に、回答可能かどうかといった不確実性の認識の失敗によるものである。現在この課題に焦点を当てた RAG の評価指標は少ない。そこで本研究では、RAG におけるテキスト生成モデルが回答可能かに関する認識の正確さを表す評価指標を自己認識的不確実性と定義し、LLM を用いてこれを評価する。

## 2 RAG

RAG の概要図を図 1 に示す。RAG は検索を行う Retriever とテキスト生成を行う Generator からなる構造をとる。まず Retriever が外部知識を検索してユーザーの質問に関連する文書を得る。カスタマーサポートの場合は外部知識として製品やサービスに関する社内のドキュメントやマニュアル、よくある質問集 (FAQ) を用いることが多い。これらは PDF や Excel といった多様な形式で保持されており、事前に検索手法に適した形式に変換する必要がある。単純なベクトル検索の場合は事前に該当するファイルから抽出したテキストを単体で意味を表す文書に区切り、それぞれに対応するベクトル表現をベクトルデータベースに格納する。検索時はユーザー

の質問に対するベクトル表現を取得し、ベクトルデータベースから類似度が高い文書を得る。そして Generator では、Retriever が検索して得た文書に基づき LLM を用いて質問に対する回答を生成する。

RAG の普及に伴い、RAG の評価方法に関する研究も盛んに行われている [5, 6]。特に RAGAS[7] では、RAG の性能を複数の観点から LLM を用いて評価する。本研究では特に RAG の Generator に関連する指標として、RAGAS の Faithfulness と Answer Relevancy に着目する。Faithfulness は事実に基づいて回答しているかを評価する指標であり、質問と回答から主張を複数生成し、それらが検索で得られた文書に基づいているかを判定する。一方、Answer Relevancy は回答がどの程度質問に関連しているかを評価する指標であり、回答と検索で得られた文書から想定される質問を生成し、元の質問との類似度を算出する。これは回答に不要な情報が含まれていると低下する傾向にある。

## 3 自己認識的不確実性

### 3.1 概要

RAG の Retriever において回答に十分な情報が検索で得られなかった場合、Generator では回答不能であることを生成することが望ましい。そこで、本研究では Generator が回答可能かに関する認識の正確さを自己認識的不確実性 (Self-Aware Uncertainty) と定義する。具体的には、検索で回答に十分な情報が得られた場合には答えの提示に向けて生成するほど、得られなかった場合には答えを提示できないことを適切に生成するほど、Self-Aware Uncertainty は高い値をとる。この時、回答の正誤は問わず、あくまで答えの提示に向けた生成をしているかどうかのみ着目し、LLM を用いて判定する。判定時のプロンプトには答えの提示に向けた生成には 1 を、それ以外には 0 を生成するように記述され、プロンプトに含まれる {question} と {answer} はそれぞれユーザーからの質問と RAG の応答に置換される。事前に定義した形式で結果を取得するために生成結果が JSON 形式となるように記述する。詳細なプロンプトは付録の表 4 に示す。判定後 (1) 検索成功かつ答えの提示に向けた生成をしている場合と (2) 検索失敗かつ答えの提示以外の生成をしている場合は適切であるため 1 を、それ以外の場合は 0 をその推論に対する評点とする。そしてこれを全てのデータにお

いて平均した値を Self-Aware Uncertainty のスコアとする。

## 3.2 回答の Rewriting

Chain of Thought[8] をはじめ、LLM は段階的な推論によって多くのタスクで推論精度が向上することが知られている。そこで、自己認識的不確実性の向上を期待し、LLM で生成した回答を再度 LLM を用いて修正する Rewriting を行う。これにより、生成した回答に対して、LLM が再度客観的に評価して修正することが期待される。このとき LLM に与えるプロンプトには {question} と {answer} と {context} が含まれており、それぞれユーザーからの質問と RAG の応答、検索で得た文書に置換される。もし回答の修正が不要であれば “VALID” という固定文字列を出力し、回答の修正が必要であれば修正した回答を出力するような指示が記述されている。Rewriting のプロンプトの詳細は付録の表 5 に示す。

## 4 実験

本研究では、公開されている Ameba FAQ データセット<sup>1)</sup>を用いて評価する。自己認識的不確実性は RAG の Generator に関する評価指標であるため、Generator のみを評価対象とする。具体的には、事前にデータセット中の質問に対して Retriever による検索を行い、その結果を用いて Generator の評価実験を行う。

### 4.1 実験設定

RAG の Retriever では、OpenAI の Embedding モデルである text-embedding-ada-002 でテキストのベクトル化を行い、コサイン類似度が高い文書を取得する。付録 B の事前実験の結果より、質問と回答の組からなる FAQ はそれらを改行区切りで連結したテキストとして扱う。そして、事前に全てのテキストに対してベクトル化を行い、ベクトルデータベースを構築する。Generator では LLM として OpenAI の GPT-3.5-turbo-16k, GPT-4-turbo, Google の Gemini-Pro[9] を採用する。生成時は出力のランダム性を制御するパラメータである Temperature を 0 に設定して回答を生成する。LLM に与えるプロンプトにはカスタマーサポートのオペレーターとして回答するような指示が記載されている。さらに、プロ

1) [https://huggingface.co/datasets/ai-shift/ameba\\_faq\\_search](https://huggingface.co/datasets/ai-shift/ameba_faq_search)

表 1 RAG の性能評価

| Model             | w/ Rewriting | Faithfulness                | Answer Relevancy            | Self-Aware Uncertainty      | Average                     |
|-------------------|--------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| GPT-3.5-turbo-16k | -            | 0.662 (0.739, 0.584)        | 0.876 (0.876, 0.876)        | 0.510 (0.876, 0.020)        | 0.683 (0.872, 0.493)        |
| GPT-4-turbo       | -            | <b>0.769</b> (0.824, 0.714) | 0.876 (0.885, 0.866)        | 0.560 (1.000, 0.120)        | <b>0.735</b> (0.903, 0.567) |
| Gemini-Pro        | -            | 0.619 (0.725, 0.513)        | 0.851 (0.862, 0.839)        | <b>0.670</b> (0.980, 0.360) | 0.713 (0.856, 0.571)        |
| GPT-3.5-turbo-16k | ○            | 0.728 (0.781, 0.674)        | <b>0.877</b> (0.873, 0.881) | 0.510 (1.000, 0.020)        | 0.705 (0.885, 0.525)        |
| GPT-4-turbo       | ○            | 0.684 (0.768, 0.600)        | 0.873 (0.882, 0.865)        | 0.570 (1.000, 0.140)        | 0.709 (0.884, 0.535)        |
| Gemini-Pro        | ○            | 0.644 (0.786, 0.502)        | 0.853 (0.864, 0.841)        | <b>0.670</b> (0.960, 0.380) | 0.722 (0.870, 0.574)        |

括弧内は左から順に、検索が成功した事例（RS 事例）、検索が失敗した事例（RF 事例）に限り評価したスコアを表す。

ンプトには固定文字  $\{context\}$  が含まれ、Retriever の検索で得た文書に置換される。プロンプトの詳細は付録の表 3 に示す。

RAG の評価指標として、3.1 節で述べた自己認識的不確実性に加え、RAGAS の Generator に関する評価指標である Faithfulness と Answer Relevancy を用いる。そして、それら 3 つのスコアの平均値を Average として計測する。評価で利用する LLM は、可能な限り高精度なモデルであることが望ましいと考えられるため全て GPT-4-turbo を用いる。

評価に関する先行研究である LLM-as-a-judge[10] では、判定に用いる LLM は同じモデルによって生成された回答を好む傾向にあることが確認されており、self-enhancement bias として知られている。そのため、本研究において推論時に Generator として GPT-4-turbo を利用している場合ではこのバイアスの影響を受けている可能性があり、正当な比較に関する課題があることに留意されたい。

## 4.2 データセット

本研究で利用する Ameba FAQ データセットは FAQ と質問データからなる。まず FAQ は質問と答えの組が 461 件存在する。次に、質問データセットは、訓練用、検証用、評価用に分かれており、それぞれ 1313 件、792 件、837 件存在する。これらのデータは質問と正解 FAQ からなる。本研究ではパラメタの更新を伴う言語モデルの学習を行わないことから、訓練用の質問データを利用しない。そして、実験に要する費用と時間の観点から、次の方法で作成した少量の検証データと評価データを利用する。

まず、全ての検証用と評価用の質問データに対して Retriever で検索した類似度上位 5 件の関連文書を得る。この結果、検証用の質問データでは関連文書に正解 FAQ が含まれている割合は 77.8% であった。次に、関連文書中に正解 FAQ が含まれているかどうかによって事例を分類し、検索が成

功した事例（Retrieve-Success: RS）と失敗した事例（Retrieve-Failure: RF）を収集する。最後に、検証用の RS 事例と RF 事例がそれぞれ 10 件ずつからなる計 20 件を検証データとし、評価用の RS 事例と RF 事例がそれぞれ 50 件ずつからなる計 100 件を評価データとする。検証データは事例の分析に利用し、実験では評価データのスコアを比較する。

## 4.3 実験結果

実験結果を表 1 に示す。まず Rewriting を行わない場合における比較では、Faithfulness において GPT-4-turbo と GPT-3.5-turbo-16k が共に Gemini-Pro を上回った一方で、Self-Aware Uncertainty において Gemini-Pro が最も高い値となった。これより、Faithfulness では捉えられなかった側面を Self-Aware Uncertainty が捉えて評価することができており、Faithfulness では Generator が回答可能かに関する認識の正確さを捉えられない傾向にあると考えられる。特に検索が失敗した RF 事例における評価ではモデル間の Self-Aware Uncertainty の差が大きく、Gemini-Pro が最も高い値である。そのため、Gemini-Pro は回答の根拠が不十分である場合に正しく回答不能であることをユーザーに伝えることができていたといえる。

次に Rewriting を行った場合における比較では、Rewriting により Average が向上したのは、GPT-3.5-turbo-16k と Gemini-Pro の 2 つであった。3 つの評価指標においては、両モデルで Faithfulness が向上しており、Answer Relevancy と Self-Aware Uncertainty に大きな差はなかった。これより、Rewriting は回答を事実に基づくように修正する点において特定のモデルで効果がある一方で、自己認識的不確実性に対しての影響は小さいと考えられる。これは Rewriting により自己認識的不確実性が向上するという我々の仮説を否定する結果となった。

最後に 3 つの評価指標の中で Answer Relevancy は

表2 検証データにおけるRF事例

|  |
|--|
| 質問:<br>スマートフォンからアメンバー限定記事を読む手順を教えてください。  |
| 回答 (GPT-3.5-turbo-16k):<br>スマートフォンからアメンバー限定記事を読む手順は以下の通りです。<br>1. Ameba ブログアプリを開きます。<br>2. ブログのトップページに移動します。<br>3. アメンバーのブロガーのブログを選択します。<br>4. アメンバー限定記事を読みたい記事を選択します。<br>5. 記事の内容を閲覧することができます。<br>以上がスマートフォンからアメンバー限定記事を読む手順です。アメンバー限定記事を読むためには、ブロガーの方に申請し承認してもらう必要がありますので、ご注意ください。 |
| 回答 (GPT-4-turbo):<br>スマートフォンからアメンバー限定記事を読むには、まずそのブロガーのアメンバーになる必要があります。アメンバーになるためには、ブロガーに申請をし、承認される必要があります。アメンバーになった後は、ブロガーがアメンバー限定で投稿した記事を読むことができます。記事の閲覧方法についての詳細は、ドキュメントに記載されている「アメンバー限定記事を閲覧する」ページをご参照ください。   |
| 回答 (Gemini-Pro):<br>情報が見つかりませんでした。   |
| Rewriting 生成結果 (GPT-3.5-turbo-16k):VALID<br>Rewriting 生成結果 (GPT-4-turbo):**VALID**<br>Rewriting 生成結果 (Gemini-Pro): 情報が見つかりませんでした。  |
| 検索で得たFAQの質問:<br>1. アメンバー限定記事を書くには?<br>2. アメンバーについて<br>3. 記事を非公開にする方法 (スマートフォン)<br>4. 携帯で書こう<br>5. 記事の編集はどこからできますか  |
| 正解のFAQの質問:<br>アメンバーってなに?   |
| {Faithfulness, Answer Relevancy, Self-Aware Uncertainty}<br>回答 (GPT-3.5-turbo-16k) に対する評価: {0.000, 0.867, 0}<br>回答 (GPT-4-turbo) に対する評価: {1.000, 0.970, 0}<br>回答 (Gemini-Pro) に対する評価: {0.000, 0.821, 1}  |

手法間で差がついていない。本研究ではRAGに与える質問はAmebaのカスタマーサポートという1つのドメインに限られた内容である。これより、同一ドメインの質問のみ与えられる実験においては、質問に対して適切に回答されているかを評価するためにはAnswer Relevancyでは不十分であると考えられる。

#### 4.4 事例分析

検証データに対する生成結果に基づいて事例の分析を行う。全体の傾向として、Gemini-Proは回答に必要な最小限のテキストを生成している一方で、GPT-3.5-turbo-16kとGPT-4-turboは「手順は以下の

通りです」といったようなカスタマーサポートのオペレーターと類似した応答文を生成している。そのため、チャットボットに与える役割によって適切なLLMが異なると考えられる。この一例を付録の表7に記載する。

次に、表2の検索が失敗したRF事例では、質問に回答するための十分な情報が検索で得られていないにも関わらず、GPT-3.5-turbo-16kとGPT-4-turboにおいて回答を生成してしまっており、hallucinationを引き起こしている。GPT-4-turboの回答に着目すると検索で得た情報には含まれていない内容が生成されており、それに対して誤ってFaithfulnessに高い値を出力している。一方で、Gemini-Proは適切に回答不能と認識しており、「情報が見つかりませんでした」と回答している。Self-Aware Uncertaintyはこれらを適切に捉えてGemini-Proのみ1となっており、Generatorが回答可能かに関する認識の正確さを、この事例においては正しく評価できているといえる。

## 5 おわりに

本研究では、RAGが回答可能かどうかの認識に関する評価指標として自己認識的不確実性(Self-Aware Uncertainty)を定めた。これはRAGの評価に関する先行研究であるRAGASの評価指標と異なり、回答可能かに関する認識の正確さを捉えている。また、LLMの段階的な推論による性能向上を期待して回答のRewritingを行った結果、GPT-3.5-turbo-16kとGemini-ProにおいてFaithfulnessの向上が確認されたが、自己認識的不確実性の向上は確認されなかった。

本研究では、AmebaのFAQという特定のドメインに限ったFAQと質問を用いて評価実験を行った。これを踏まえ、今後はチャットボット事業を通して収集した複数のドメインにおけるRAGの性能評価を実施する。また、今回の実験では費用と時間の観点から小規模なデータセットを用いた。今後はより大規模なデータを用いて実験を行い、詳細な事例分析と共に自動評価と人手評価との相関を計測し、適切な評価指標の策定を行う。さらに、カスタマーサポート事業への応用においてより優先的に評価すべき観点を事例を通して分析することで、評価に費やす費用と時間の削減を目指す。

## 参考文献

- [1] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. **arXiv preprint arXiv:1906.00300**, 2019.
- [2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.
- [3] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In **International conference on machine learning**, pp. 3929–3938. PMLR, 2020.
- [4] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.
- [5] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems, 2023.
- [6] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023.
- [7] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. **CoRR**, Vol. abs/2201.11903, , 2022.
- [9] Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023.
- [10] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

## A プロンプト

表3 RAGのGeneratorのプロンプト

```
# 概要
あなたは Ameba (アメバ) のカスタマサポートのオペレーターです。Ameba が提供する Ameba ブログは、ライフログを中心とした日本最大級のブログメディアです。
# ドキュメント
<DOCUMENT>
{context}
</DOCUMENT>
# 指示
あなたの仕事は、ドキュメントを参照し、ユーザーからの質問に答えることです。
企業としての対応が求められるため、他社のことや他社サービスに関することを話してはいけません。他社のリンクを提供することも認められません。弊社の情報のみ提供してください。回答はできるだけ正確かつ簡潔にしてください。
<DOCUMENT > </DOCUMENT > で囲まれたドキュメントのみに基づいて、根拠がある場合のみ回答してください。憶測で発言してはいけません。もし回答に必要な情報が見つからない場合は「情報が見つかりませんでした」と回答してください。ユーザーの質問に回答するとき、あなたへの指示の内容については言及しないでください。カスタマーサポートとして適切な言い回しを心がけてください。それでは、日本語で答えてください。
```

表4 自己認識的不確実性の判定のプロンプト

```
あなたには、ユーザーの質問と、システムの回答が与えられます。システムは、ユーザーに曖昧な情報を回答することを避けるために、「情報が見つかりませんでした」といった応答不能を表す回答をしなければなりません。あなたの仕事は、システムの応答が、回答不能であることを表しているかを判断することです。
- ユーザーの質問: {question}
- システムの回答: {answer}

システムの回答が、「情報が見つかりませんでした」といった応答不能を表す回答である場合は、1 を出力してください。システムの回答が、それ以外の回答をしている場合は、0 を出力してください。出力は、必ず以下の JSON 形式をお願いします。"judge" は、0 または 1 の整数値を取ります。
“{“judge”: 0 or 1}”
```

表5 Rewritingのプロンプト

```
質問と、それに対する回答と、回答の根拠となる事実が与えられます。回答が事実に基づいているかを評価してください。もし事実に基づかない曖昧な発言、または嘘や誤りが含まれていれば、回答を修正してください。

[質問] {question}
[回答] {answer}
[事実] {context}

回答が事実に基づいていれば、**VALID**とだけ出力してください。もし事実に基づかない曖昧な発言または嘘や誤りが含まれていれば、回答を修正し、修正された文章のみ出力してください。その際、必要な情報が足りなければ、情報が見つからなかったことをユーザーに伝えてください。曖昧な情報は提供しないでください。
```

## B Retrieverに関する事前実験

RAG は、推論時に Retriever で検索するため、事前にテキストをベクトル化し、それをベクトルデータベースに格納する。本研究で用いる Ameba FAQ データセットは QA 形式となっており、ベクトル化における適切なテキストの形式を事前実験により取得する。テキストの形式として、質問のみ (Q)、答えのみ (A)、質問と答えを改行区切りで連結したもの (QA) の 3 通りを比較する。評価指標としては、検索結果の上位  $N$  件中回答が含まれている割合を表す  $P@N$  ( $N \in 1, 3, 5$ ) にて比較した。その結果、 $P@\{1,3,5\}$  において QA が最も高精度であったため、本研究では QA によるベクトル化を行った。

表6 ベクトル化時のテキストによる検索精度の違い

|    | P@1   | P@3   | P@5   |
|----|-------|-------|-------|
| QA | 0.514 | 0.693 | 0.778 |
| Q  | 0.367 | 0.497 | 0.558 |
| A  | 0.435 | 0.621 | 0.717 |

## C 検証データ中の生成事例

検証データの内、RAG の検索が失敗した RF 事例を表 2 に示した。ここでは、RAG の検索が成功した RS 事例について表 2 に示す。3 つのモデルの回答はいずれも適切であり、Faithfulness, Answer Relevancy, Self-Aware Uncertainty の全てが高い値である。しかし、Rewriting 生成結果において、Gemini-Pro では回答の修正が行われている。さらに、修正された回答のみだけでなく、修正に伴う説明を生成している。これより、Rewriting では LLM の出力の制御性に伴う影響を受けると考えられる。

表7 検証データにおける RS 事例

```
質問:
新しいメールアドレスを登録するためには、どのような手続きが必要ですか?
回答 (GPT-3.5-turbo-16k):
新しいメールアドレスを登録するためには、以下の手順をお試しください。
1. Ameba の会員登録情報ページにアクセスしてください。
2. 登録メールアドレス項目の「メールアドレスを変更する」をクリックしてください。
3. 新しいメールアドレスを入力し、「変更」をクリックしてください。
4. 入力したメールアドレス宛に「【Ameba】メールアドレス変更のお知らせ」というタイトルのメールが届きます。
5. メール本文内の URL をタップまたはクリックし、変更を完了してください。
なお、すでに他のアカウントに登録されているメールアドレスへの変更を行うとすると、入力されたメールアドレスに誤りがあると表示され変更ができません。登録のないメールアドレスへの変更をお願いいたします。また、変更後のメールアドレスに関連するメールは、変更前のメールアドレス宛にも一部送信される場合がありますのでご了承ください。
回答 (GPT-4-turbo):
新しいメールアドレスを登録するための手順は以下の通りです。
1. 会員登録情報ページへアクセスしてください。
2. 登録メールアドレス項目の「メールアドレスを変更する」をクリックしてください。
3. ご希望の新しいメールアドレスを入力し、「変更」をクリックしてください。
4. 入力した新しいメールアドレス宛に「【Ameba】メールアドレス変更のお知らせ」というタイトルのメールが届きます。
5. メール本文内の URL をタップまたはクリックし、変更を完了してください。
なお、すでに他のアカウントに登録されているメールアドレスへの変更はできませんので、登録のないメールアドレスをご確認ください。
回答 (Gemini-Pro):
1. 会員登録情報ページにアクセスしてください。 ※スマートからのみご利用の場合はこちらのページへアクセスして変更を行ってください。
2. 登録メールアドレス項目の【メールアドレスを変更する】をクリックしてください。
3. ご希望のメールアドレスを入力し、【変更】をクリックしてください。
4. 入力したメールアドレス宛に「【Ameba】メールアドレス変更のお知らせ」というタイトルのメールが届きます。
5. メール本文内の URL をタップ・クリックし、変更を完了させてください。
Rewriting 生成結果 (GPT-3.5-turbo-16k):VALID
Rewriting 生成結果 (GPT-4-turbo):**VALID**
Rewriting 生成結果 (Gemini-Pro): 回答が事実に基づいていません。修正した回答は、以下のとおりです。
ご希望のメールアドレスを入力していただき、パスワードの再設定をお試しください。パスワードを忘れた場合より新規登録を行いたいメールアドレスを入力していただき、パスワードの再設定をお試しください。
→
申し訳ございませんが、ご希望のメールアドレスがすでに登録されているため、パスワードの再設定はできません。Ameba では、1 つのアカウントにつき 1 つのメールアドレスを登録することができます。そのため、新しいメールアドレスを登録するためには、新しいアカウントを作成する必要があります。
検索で得た FAQ の質問:
1. メールアドレスが既に登録されていますと表示されます (正解の FAQ)
2. 登録したメールアドレスは変更できますか?
3. 入力されたメールアドレスに問題がありますと表示されます
4. 登録したメールアドレスを変更したいのですが
5. メールアドレスの変更について
{Faithfulness, Answer Relevancy, Self-Aware Uncertainty}
回答 (GPT-3.5-turbo-16k) に対する評価: {1.000, 0.804, 1}
回答 (GPT-4-turbo) に対する評価: {1.000, 0.894, 1}
回答 (Gemini-Pro) に対する評価: {1.000, 0.804, 1}
```